Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 38 (2012) 1501 – 1507

Procedia Engineering

www.elsevier.com/locate/procedia

Efficient Statistical Clustering Techniques for Optimizing Cluster Size in Wireless Sensor Network

Raju Dutta[a], Shishir Gupta[b], Mukul K. Das[a,b,a]*

[a]*Asst. Professor, Dept. of Mathematics, Narula Institute of Technology, Kolkata 700109,West Bengal, India*
[b]*Professor , Dept. of Mathematics, Indian School of Mines, Dhanbad 826004, Jharkhand,India*

Abstract

Mobility of sensor node in Wireless Sensor Network (WSN) is one of the key advantages of wireless over fixed communication system. In heterogeneous system, generally power consumption is more then homogeneous system. The information or data message passing process must be well architecture to save the limited energy resources of the sensors. Clustering of sensors into different groups, so that sensors communicate information to the other cluster and then the cluster communicate the true information to the processing center which may save energy. So the **c**oordination in distributed sensor network the implementation of clustering is an important technique and clusters of bounded size which is the total number of nodes in a specific cluster, is an important parameter in clustering algorithms which are very much effective in reducing energy consumption by minimizing the neighborhood of a node. Communication cost is also an important parameter for computation in a large area. Clustering techniques is Wireless Sensor Networks (WSNs) compare to random selection techniques is less costly due to the saving of time in journeys, reduction in number of transmissions and receptions at each node, identification, contacts etc. Which are valuable for increasing the overall network life, scalability of WSNs. Clustering sensor nodes is an effective and efficient technique for achieving all the requirement**.** In this paper, we propose a distributed, randomized clustering techniques to find optimum cluster size and cost to organize the sensors in a wireless sensor network within clusters.

Keywords **—** sensor networks, clustering, cluster size, energy efficiency, network lifetime.

1. INTRODUCTION

A wireless sensor network forms by self-organized sensors and many of the node are not perform well all the time in the network. Every node has a specific application in an area in the network. In random sampling it is assumed that the sensor population has been divided into a finite number of distinct and identifiable sampling units called element or sensor of the population. Generally, identification and location of an element requires considerable time. However, once an element has been located, the time taken for surveying a few neighbouring elements are small. Thus the main function in cluster sampling is to specify clusters or to divide the population into appropriate clustering, therefore, the elements showing

doi:10.1016/j.proeng.2012.06.185

similar characteristics within a cluster. As a simple rule, system will perform actively if the number of elements in a cluster is small and the number of clusters should be large. The required number of clusters can be selected either by equal or unequal probabilities of selection after dividing the population into specified clusters. The efficiency of cluster sampling is likely to decrease with increase in cluster size. So for a given sample size, a smaller sampling unit will bring more precise results then a larger sampling units. For a given number of sampling units, cluster sampling is more convenient and less costly. The advantages of cluster sampling are collection of data for neighbouring elements is easier, cheaper, faster and operationally more convenient then observing units spread over a region. It is less costly then simple random sampling due to the saving of time in journeys, identification, contacts etc. When a sampling frame of elements may not be readily available.

1.1. Related Work

Wireless Sensor Networks (WSNs) have recently became an interesting field which emerged as an important computing platform, discussed in[1], [2]. In [3] Clustering techniques can give an overview in reducing useful energy consumption. Clustering is very much useful for applications that require scalability to hundreds or thousands of nodes. Scalability in this context implies the need for load balancing and efficient resource utilization. Applications requiring efficient data aggregation (e.g., computing the maximum detected radiation around an object) are natural candidates for clustering. Routing protocols can also employ clustering [4], [5]. In [6], clustering was proposed as a useful tool for efficiently pinpointing object locations. Clustering can be extremely effective in one-to-many, many to-one, one-to-any, or one-to-all (broadcast) communication. This paper [7] uses clusters to transmit processed data to base stations, hence minimizing the number of nodes that take part in long distance communication. This directly affects the overall system energy dissipation. In [8], describes clustering for efficient standard cell placement. Clustering algorithms [9-13] separate a network into clusters and each cluster has a clusterhead (CH) that acts as a local controller. The efficiency of cluster sampling has been studied by Smith (1938), Hansen and Hurwitz (1942). Mahalanobis (1940, 1942) have studied the problem of determination of the optimum cluster size from the point of view of both variance and cost. A comprehensive study of cluster size in the cluster sampling and sub-sampling procedure has been made by Singh (1956). In this paper we discuss the problem of optimum size of cluster for which maximum precision is attained with a given cost, or vice versa. We propose a strategy that differs from the other strategies in various aspects.

1.2. System Mathematical Notations

We shall first consider the case of equal clusters. Suppose the population consists of N clusters, each of M elements and that a sample of n clusters is drawn by method of simple random sampling. y_{ij} = the value of the characteristics under study for the j^{th} element, (j =1,2,……..,M) in the i^{th} cluster, (i =1,2,……..N)

$\bar{y}_{i.} = \sum_{j}^{M} \frac{y_{ij}}{M}$ = the mean of per element of the ith cluster

$\bar{y}_{n} = \sum_{i}^{n} \frac{\bar{y}_{i.}}{n}$ = the mean of cluster means in the sample of n clusters

$\bar{Y}_{N} = \sum_{i}^{N} \frac{\bar{y}_{i.}}{N}$ = the mean of cluster means in the population

$S_i^2 = \sum_{j}^{M} \frac{\left(y_{ij} - \bar{y}_{i.}\right)^2}{\left(M - 1\right)}$ =the mean square between elements within the i^{th} cluster (i = 1,2,……..,N)

$S_w^2 = \sum_i^N \frac{S_i^2}{N}$ = the mean square within the cluster.

$S_b^2 = \sum_i^N \frac{\left(\bar{y}_{i.} - \bar{Y}_N\right)^2}{(N-1)}$ = the mean square between clusters means in the population.

$S^2 = \sum_i^N \sum_j^M \frac{\left(y_{ij.} - \bar{Y}\right)^2}{(NM-1)}$ = the mean square between elements in the population

$\rho = \frac{E\left(y_{ij} - \bar{Y}\right)\left(y_{ik} - \bar{Y}\right)}{E\left(y_{ij} - \bar{Y}\right)^2}$ = the intracluster correlation coefficient between elements within clusters.

1.3. Problem Definition and analysis

EQUAL CLUSTER SAMPLING

No new principles are involved in making estimates when a probability sample of "n" equal sized clustering has been considered in WSNs and each cluster is enumerated completely. Since the cluster is of equal size, it is clear that $\bar{Y}_N = \bar{Y}$. For the sampling variance of the estimator $\bar{y}_n$. we shall begin by the following consideration.

Estimator of Mean and its Variance

In Simple Random Sampling (SRS), Without Replacement (WOR), of n clusters each containing M elements from population of N clusters, the sample mean $\bar{y}_n$ is an unbiased estimator of $\bar{Y}$ and its variance is given by

$$V\left(\bar{y}_n\right) = \frac{(1-f)}{n} S_b^2 \tag{1}$$

where $f = \frac{n}{N}$ is called sampling fraction

on simplifying by substituting S_b^2 and putting value of the intracluster correlation coefficient ρ, we get

$$V\left(\bar{y}_n\right) = \frac{(1-f)}{nM} S^2 \left[1 + (M-1)\rho\right] \tag{2}$$

for large N.

The factor $\left[1+(M-1)\rho\right]$ shows that how much the variance is changed by the use of a cluster instead of elements as sampling unit.

It has been shown that the variance in cluster sampling depends on the number of clusters in the sample, the size of the cluster, the intracluster correlation coefficient ρ and the variance S^2. If M=1 it gives the sampling variance of a SRS of nM elements are taken individually and which are equally good in the cluster sampling. If M>1 and $\rho \rangle 0$, is positive, cluster sampling will give a higher variance then the mean per element then the cluster is less precise for a given sample. If ρ is negative then the cluster sampling is more precise that is the main reason to use cluster sampling.

Corollary 1: In SRS, wor, of n cluster each containing M elements from a population of N clusters the population total is estimated by

$$\bar{Y}_C = N M \bar{y}_n$$

$$V\left(\bar{Y}_C\right) = (N M)^2 V\left(\bar{y}_n\right) \cong M N^2 \frac{(1-f)}{n} S^2 \left[1 + (M-1)\rho\right] \tag{3}$$

1.4. Relative Efficiency of Equal Cluster Sampling

In sampling nM elements from the population by SRS, the variance of the sample mean $\bar{y}$ is given by

$$V(\bar{y}) = \frac{(1-f)}{nM} S^2 \tag{4}$$

Thus, the relative efficiency of cluster sampling compared with SRS is given by

$$\text{Relative Efficiency} = \frac{V_{SR}(\bar{y})}{V_C(\bar{y}_n)} = \frac{S^2}{M S_b^2} \tag{5}$$

This shows that the efficiency of cluster sampling increases as the mean square between clusters decreases and the relation between S_b^2, S^2 and S_W^2 is given below in the following expression

$$(N-1)MS_b^2 = (NM-1)S^2 - N(M-1)S_W^2 \tag{6}$$

Therefore, relative efficiency will increase with increase in the mean square within clusters. These results suggests that the clusters should be so formed that variance within clusters is maximum. Another way to express relative efficiency is to take use of the concept of the intracluster correlation coefficient.

For large N, the relative efficiency of clusters sampling in terms of intracluster coefficient ρ is given by

$$\text{relative efficiency (E)} = [1+(M-1)\rho]^{-1} \tag{7}$$

Case 1: in the case of complete homogeneity of clusters, $S_w^2 = 0$ and so $\rho = 1$ and $E = \frac{1}{M}$ i.e cluster sampling is not efficient.

Case 2: in the case of complete heterogeneity of clusters, $S_w^2 = S^2$ and so $S_b^2 = 0$ and $\rho = -\frac{1}{M-1}$ i.e cluster sampling is very much efficient.

Hence, it should be noted that ρ lies in the range $-\frac{1}{(M-1)} \le \rho \le 1$. It also shows that cluster sampling will be more efficient if ρ is negative. In practice, ρ is usually positive as neighbouring elements are grouped to form clusters. Generally, ρ decreases with increase in M. The efficiency of cluster sampling increases as the factor $[(M-1)\rho]$ increase with cluster size.

The efficiency can easily be calculated by estimating the value of ρ from the sample. An estimator of ρ can be written as

$$\hat{\rho} = \frac{(n-1)MS_b^2 - nS_w^2}{(n-1)MS_b^2 + n(M-1)S_w^2} \quad \text{where} \quad s_w^2 = \frac{\sum_i \sum_j^{nM} (y_{ij} - \bar{y}_{i.})^2}{n(M-1)} \tag{8}$$

Thus for large N, an estimator of the relative efficiency of cluster sampling can be written as

$$\text{Est. Relative Efficiency (e)} = \frac{1}{M} + \frac{(M-1)s_w^2}{M^2 s_b^2} \tag{9}$$

and accordingly ρ can be estimates by $\hat{\rho} = \frac{(1-e)}{(M-1)e}$ (10)

It should be noted clearly that in a random sampling of n clusters, s_b^2 and s_w^2 will provide unbiased estimator of S_b^2 and S_w^2. Where s^2 will not be an unbiased estimator of S^2. The reason is that a sample of nM elements is not taken randomly from the population of NM elements.

However, an unbiased estimator may be obtained easily by substituting the values in the relation, where

$$\hat{S}^2 = \frac{(N-1)M s_b^2 + N(M-1)s_w^2}{(NM-1)} \tag{11}$$

1.5. Optimum Cluster Size

For a given sample size, the sampling variance increase with cluster size and decrease with increasing number of clusters. On the other hand, the cost decrease with the cluster size and increase with the number of clusters. Hence, it is necessary to determine a balancing point by finding out the optimum cluster size and the number of clusters in the samples which can minimize the sample variance for a given cost or, alternatively, minimize the cost for a fixed variance.

Here we have assumed the cost of a survey, apart from overhead cost, will be made up of two components:

(i) Cost due to expanses in enumerating the elements in the sample and in traveling within the cluster, which is proportional to the number of elements in the sample.

(ii) Cost due to expanses on traveling between clusters, which is proportional to the distance to be traveled between clusters. It has been shown empirically that the expected value of minimum distance between n points located at random is proportional to $n^{\frac{1}{2}}$.

The cost of the survey can be, therefore, expressed as $C = c_1 n M + c_2 n^{\frac{1}{2}}$ (12)

where c_1 is the cost of enumerating an element, including the cost of travel between units within the cluster, and c_2 is the cost per unit distance traveled between clusters.

It has already been shown that the variance of the estimator $\bar{y}_n$ based on a sample of n clusters of size M each, is given by $V(\bar{y}_n) = \frac{(1-f)}{n} S_b^2$ (13)

As shown by relation (6) S_b^2 can be obtained if we know

(i) the variance S^2 between all elements in the population and (ii)The variance S_w^2 within clusters.

Hence, an approach has always been made to predict S_w^2 as it is affected by the cluster size while S^2 remains unchanged by it. On the basis of several agricultural surveys, it has been observed that the energy dissipation to transmit a massage to a specific distance estimated by the relation with M which can be written as by empirical relation [14] $S_w^2 = a M^b \quad (b > 0)$ (14)

where a and b are positive constant for a specific wireless system (usually $2 < b < 4$) and are to be determined from the survey data and do not depends on M.

From the analysis of variance, we have

$$S_b^2 = \frac{(MN-1)S^2 - N(M-1)S_W^2}{M(N-1)} \tag{15}$$

By substituting the value of from relation (14), for large N we have

$$S_b^2 = S^2 - (M-1)a M^{b-1} \tag{16}$$

In relation (13), after ignoring the finite population corrections (fpc) and substituting the value of S_b^2 from relation (16) we have $V(\bar{y}_n) = \frac{1}{n}[S^2 - (M-1)a M^{b-1}]$ (17)

First solve the cost equation (12) as a quadratic equation in $n^{1/2}$. This gives

$$\frac{2c_1 M\sqrt{n}}{c_2} = \left(1 + \frac{4Cc_1 M}{c_2^2}\right)^{\frac{1}{2}} - 1 \tag{18}$$

On simplifying the above expression we thus obtained the value of n is given by

$$n = \left[\frac{-c_2 + \left(c_2^2 + 4c_1CM\right)^{\frac{1}{2}}}{2c_1M} \right]^2 \tag{19}$$

Now by the manipulation we can get the expression that gives the optimum M. The problem under consideration is to calculate the value of n and M by minimizing V for given fixed C. To minimize $\phi = C + \lambda V$ (λ Being constant multiplier),

$$\phi = C + \lambda V = c_1 nM + c_2 n^{\frac{1}{2}} + \lambda V .$$

Differentiating w.r.to n and M, and noting that $\frac{\partial V}{\partial n} = -\frac{V}{n}$ we obtain the equation

$$c_1 M + \frac{1}{2} c_2 n^{-\frac{1}{2}} = -\lambda \frac{\partial V}{\partial n} = -\frac{V}{n} \text{ and } c_1 n = -\lambda \frac{\partial V}{\partial M}$$

Eliminating λ from the above relation we get $\frac{n}{V}\frac{\partial v}{\partial M} = \frac{c_1 n}{c_1 M + \frac{1}{2} c_2 n^{-\frac{1}{2}}} = \frac{1}{1 + \frac{c_2}{2c_1 M\sqrt{n}}}$

If we substitute for $\sqrt{n}$ from ( ) we obtain after simplification

$$\frac{M}{V}\frac{\partial V}{\partial M} = \left[\left(1 + \frac{4Cc_1 M}{c_2^2} \right)^{-\frac{1}{2}} - 1 \right] \tag{20}$$

now differentiating (17) w.r.to M and Using in (19)get an explicit expression for M. However, M can be obtained from the equation $\frac{aM^{b-1}\left[bM - (b-1)\right]}{S^2 - (b-1)a M^{b-1}} = 1 - \left(1 + \frac{4c_1 CM}{c_2^2} \right)^{-\frac{1}{2}}$ (21)

By iterative method. On substituting the value of M thus obtained in relation (19), we can obtain the optimum value of n.

1.6. Conclusion

c_1 contains the cost of the interaction and the cost of travel from unit to unit within cluster, these facts lead to the conclusion that the optimum size of unit becomes smaller when c_1 increases if the length of interview increases, whereas c_2 decreases if travel becomes cheaper or if the units in a given area becomes denser and total amount of money (C) used increases. This conclusion is a consequence of the type of cost function and would require reexamination with a different function. It illustrates the fact that the optimum unit is not a fixed characteristic of the population, but depends also on the type of survey and on the levels of prices and wages. After comprehensive study we conclude that (3) Implies the variance of population mean is greater then variance of sample mean. The relative efficiency of equal clustering compared with SRS in sensor network is given in (5) shows that the efficiency of clustering increases as the mean square between clusters decreases. The optimal cluster size "n" will be obtain form the expression (19). Apart from the cost function has been derived in (12) which will be optimize if the cost of enumerating an element, including the cost of travel between units within the cluster, and the cost per unit distance traveled between clusters are optimum. Hence, it is observed that the efficiency increase as the variation between clusters decreases.

References

[1] D. Estrin, L. Girod, G. Pottie, and M. Srivastava, "Instrumenting the World with Wireless Sensor Networks," in *International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2001), Salt Lake City, Utah*, May 2001. [Online]. Available: http://citeseer.nj.nec.com/estrin01instrumenting.html.

[2] G. J. Pottie and W. J. Kaiser, "Wireless Integrated Newtork Sensors," *Communications of the ACM*, vol. 43, no. 5, pp. 51–58, May 2000.

[3] W. R. Heinzelman, A. Chandrakasan, and H. Balakrishnan, "An Application-Specific Protocol Architecture for Wireless Microsensor Networks," *IEEE Transactions on Wireless Communications*, vol. 1, no. 4, pp. 660–670, October 2002.

[4] C. R. Lin and M. Gerla, "Adaptive Clustering for Mobile Wireless Networks," in *IEEE J. Select. Areas Commun.*, September 1997.

[5] S. Banerjee and S. Khuller, "A Clustering Scheme for Hierarchical Control in Multi-hop Wireless Networks," in *Proceedings of IEEE INFOCOM*, April 2001.

[6] D. Estrin, R. Govindan, J. Heidemann, and S. Kumar, "Next Century Challenges: Scalable Coordination in Sensor Networks," in *Proceedings of the ACM/IEEE International Conference on Mobile Computing and Networking (MOBICOM)*, August 1999.

[7] Heinzelman, W. R.; Chandrakasan, A.; Balakrishnan, H. Energy efficient Communication Protocols for Wireless Microsensor Networks, *Proc. Hawaaian Int'l Conf. on Systems Science*, 2000.

[8] Kleinhans, J. M.; Sigl, G.; Johannes, F. M.; Antreich, K. J. GORDIAN: VLSI Placement by Quardratic Programming and Slicing Optimization, *IEEE Trans on Computer Aided Design* 1991.

[9] W. R. Heinzelman, et al., "Energy-Efficient Communication Protocol for Wireless Microsensor Networks," Proc. of 33rd IEEE Hawaii International Conference on System Science, Jan. 2000.

[10]J. S. Liu and C. H. Richard Lin, "Energy-Efficiency Clustering Protocol in Wireless Sensor Networks," Ad Hoc Networks, 2005, (3): 371-388.

[11] O. Younis and S. Fahmy, "HEED: A Hybrid, Energy-Efficient, Distributed Clustering Approach for Ad Hoc Sensor Networks," IEEE Transactions on Mobile Computing, 2004, 3(4): 366-379.

[12] S. D. Muruganathan, D. C. F. Ma, R. I. Bhasin, and A. O. Fapojuwo, "A Centralized Energy-Efficient Routing Protocol for Wireless Sensor Networks," IEEE Radio Communications, March 2005, pp S8-S13.

[13] Mao Ye, Chengfa Li, Guihai Chen, and Jie Wu, "EECS: An Energy Efficient Clustering Scheme in Wireless Sensor Networks," Proc. IPCCC 2005, pp. 535-540.

[14] Dali Wei, Chan H.A., "A Survey on Cluster Schemes in Ad Hoc Wireless Networks," 2nd International Conference on Mobile Technology, Applications and Systems, 2005, pp. 1 - 8